

Augmenting Electronic Environments for Leadership¹

Joseph Psotka

Ken Robinson

US Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Lynn Streeter

Thomas Landauer

Karen Lochbaum

Knowledge Analysis Technologies
Boulder, CO 80301

Summary

Adapting instructional content to match the background knowledge of the student has been a long-standing goal of student modeling and tutoring (Bruner, 1966; Burton & Brown, 1979). To this end, cognitive scientists have developed student models that rely on manual knowledge bases relevant to the particular instructional task at hand. Normally, these include domain content, instructional content, models of student misconceptions, and more. While tailoring instruction to the learner has been shown to be effective (Anderson, 2002), current approaches are difficult to implement because of the enormous amount of skilled professional effort required.

Ideally, the system should *automatically* select the most appropriate content for the student based on a minimal amount of student data. This educational desideratum has been coined the “Goldilocks Principle”—providing lessons, texts, probes, etc. that are neither too advanced nor too elementary, but just right—in the “zone of proximal development (ZPD)” (Palincsar & Brown, 1984; Vygotsky, 1978), known to be important for promoting efficient learning.

In an early study, Wolfe et al. (1998) used Latent Semantic Analysis (LSA) to automatically select the next text passage for students to read and achieved one-sigma learning augmentation effects (Bloom, 1976). In the study reported here, we examined the Goldilocks Principle in the context of an LSA - enhanced online discussion environment, where contributions were automatically selected to be most similar in meaning to learners’ notes or contributions. We found that these contributions were almost always of somewhat higher judged quality. We suggest that this implementation of the Goldilocks Principle is a consequence of how LSA represents consensual knowledge, and provides an automatic way for selecting the next best piece of material for a student to learn, making this an important contribution to tutoring.

INTRODUCTION

Student Modeling and Tutoring using Latent Semantic Analysis

Automatically adapting the content of instruction to students’ background knowledge has long been the holy grail of student modeling and tutoring (Bruner, 1966; Burton & Brown, 1979). Four types of cognitive models have been developed for use in interactive intelligent tutors: global descriptions in terms of components such as cognitive and learning styles; overlay or tracing models specifically describing the micro units of instruction; models of student errors and misconceptions; and models that structure the knowledge to be learned at graduated depths of expertise (Psotka, Massey, and Mutter, 1988). The Achilles’ heel of all these approaches is the large cost of developing the models. As effective as they are, they are very difficult to implement, largely because of the enormous professional efforts required to organize the various knowledge bases for subject matter content, instructional strategies, student assessment of bugs and misconceptions, and knowledge levels. To move student modeling to a more practical plane, automated

¹ This paper does not represent US Army official policy. Approved for public release. Please contact the lead author, Dr. Joseph Psotka, US Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Ave., Alexandria, VA 22333 at Joseph.psotka@us.army.mil.

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 00 APR 2004		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE Augmenting Electronic Environments for Leadership				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Institute 5001 Eisenhower Ave. Alexandria, VA 22333; Knowledge Analysis Technologies Boulder, CO 80301				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES See also ADM001667, NATO RTO-MP-HFM-101 Advanced Technologies for Military Training (Technologies avancées pour l'entraînement militaire)., The original document contains color images.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 16	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

methods of locating and constructing appropriate knowledge, and to code, test, and assure the effectiveness of the representation, delivery, and interaction software are needed.

Latent Semantic Analysis (LSA), a machine learning technology for simulating human meaning of words and text passages appears to be appropriate for the student modeling and tutoring problem. LSA is both a model of human knowledge representation and a method for extracting and representing the meaning of words mathematically (for greater detail, see Landauer and Dumais, 1997; Landauer et al., 1998; Landauer, 2002). LSA induces word and passage meanings by mathematically analyzing a large corpus of relevant text. In the result, every word and every passage are represented as points in a high-dimensional "semantic space." This space defines the degree of estimated semantic similarity between any two words or passages. Simulations of many linguistic, psycholinguistic, and human learning phenomena, as well as several educational applications, show that LSA very accurately reflects corresponding similarities of meaning as judged or used by humans (Landauer and Dumais, 1997; Foltz et al., 1999). As we will show, LSA also provides an expeditious approach to the creation of tutoring models, by learning how to relate the semantic similarity of many different written descriptions and written answers to complex problems, and combining and presenting them automatically in effective sequences.

Wolfe et al. (1998) applied LSA directly to the problem of finding the next best piece of text for a student to read. They reasoned that the ability of a reader to learn from text depends on the match between the background knowledge of the reader and the difficulty of the text information. They used LSA to measure the distance between the reader's knowledge as gleaned from their short essays and graduated essays on the structure and function of the heart, collected from four sources of increasing difficulty.

In their study, college and medical students wrote short essays on the anatomy and function of the heart. The students next read one of four texts that ranged in difficulty from elementary to medical school level, and then wrote a new essay. Results showed that learning was greatest for texts that were neither too easy nor too difficult. Essays were represented in an LSA space which included articles on the heart and circulatory system. Degree of difficulty between the essay and a piece of text was indexed by the similarity between them as measured by the cosine in the same space. For each essay the vector in this 100 dimensional space was found and compared to the four target texts. A low cosine value between the text and essay would indicate low similarity, and thus reading this text would produce little learning. On the other hand, a too high cosine would indicate that the student already knew the content of text, and thus would be unlikely to learn anything new from the text passage. Learning, as measured by pre – post gains in short answer tests and essays (independently scored by ETS), was greatest for intermediate cosine similarity values. A more advanced version of the method (Rehder et al., 1998) placed all the essays on a line, so that cosines indicating similarity of texts that were less advanced and those that were more advanced could be distinguished. This produced even stronger results.

The LSA cosine similarity measure proved as effective at predicting learning from these texts as the traditional knowledge assessment measures. The implication is that, like Goldilocks, a student should be offered explanatory material that is neither too easy nor too difficult, but just right. Optimally the texts should stretch the student's understanding but still be comprehensible, and introduce some, but not too many new concepts. Typically in the Wolfe et al. results, the best learning occurred with texts whose cosine distances from student essays were in the range of 0.5 to 0.6. Although the magnitude of these cosines is particular to the dimensionality of each singular value decomposition semantic space, it is instructive that they lay somewhere in the middle of the positive range.

Applying Goldilocks to Collaborative Learning

In the study reported in this paper, the Goldilocks Principle was applied to contributions in a customized electronic discussion environment. If a broad range of novices, journeymen, and experts write about their

approaches to solving a problem, it should be possible to analyze their contributions systematically and automatically, using LSA, to provide graduated responses that vary in conceptual difficulty, complexity, and number of constituent themes that compose solutions. Although in this work, we were not yet able to vary the responses dynamically during discussions, we were able to analyze the spontaneous interactions in ways that serve to evaluate the hypothesis. The analyzed interactions were discussion contributions written by US Army officers who varied in rank and range of expertise. Contributions were later separately rated for overall quality by expert raters. We then examined whether the spontaneous notes of other officers, presumably ones offered on the basis of different knowledge and intent, could be used to implement the Goldilocks Principle. The goal was to find an automated way to find the next best discussion contribution or text sample for the student to read—something that was just right. As a first order approximation, the system did dynamically point the officers to other contributions that were most similar to theirs, as measured by LSA.

Problem Scenarios

Tacit Knowledge of Military Leadership (TKML)

Four different scenarios dealing with military management situations were used in the online discussion environment. The scenarios were developed by Yale University, in collaboration with the Army Research Institute to assess Tacit Knowledge of Military Leadership (TKML) (see Hedlund et al., 2000; Sternberg et al., 2000). The method is based on a carefully developed set of representative scenarios of challenging interpersonal leadership situations that are commonly encountered by Army officers, along with sets of alternative actions that a leader might take. Interviews with experienced officers originally suggested the scenarios and alternatives. Scenarios are refined, tested, and edited through a process of expert judgments, trial uses, and evaluation against a criterion of showing more expert-like responses with increasing military rank. In previous TKML administrations the leader or trainee first reads one of the scenarios, then rates on a nine-point scale the appropriateness of each of six to ten or more alternative actions that are described. This is not a multiple-choice test in the sense that only one answer is correct. The alternative actions are all acceptable or unacceptable to various degrees, the mix varying from scenario to scenario, and the “right” rating of each alternative defined by expert consensus. The test was first validated as a survey over three levels of leadership: platoon, company, and battalion command. As expected, more experienced leaders agreed on the most effective courses of action to a much greater extent (Hedlund et al., 2000) than less experienced officers, reflecting tacit leadership knowledge acquired on the job.

An example of one of the platoon-level scenarios is:

You are a new platoon leader who takes charge of your platoon when they return from a lengthy combat deployment. All members of the platoon are war veterans, but you did not serve in the conflict. In addition, you failed to graduate from Ranger School. You are concerned about building credibility with your soldiers. What should you do?

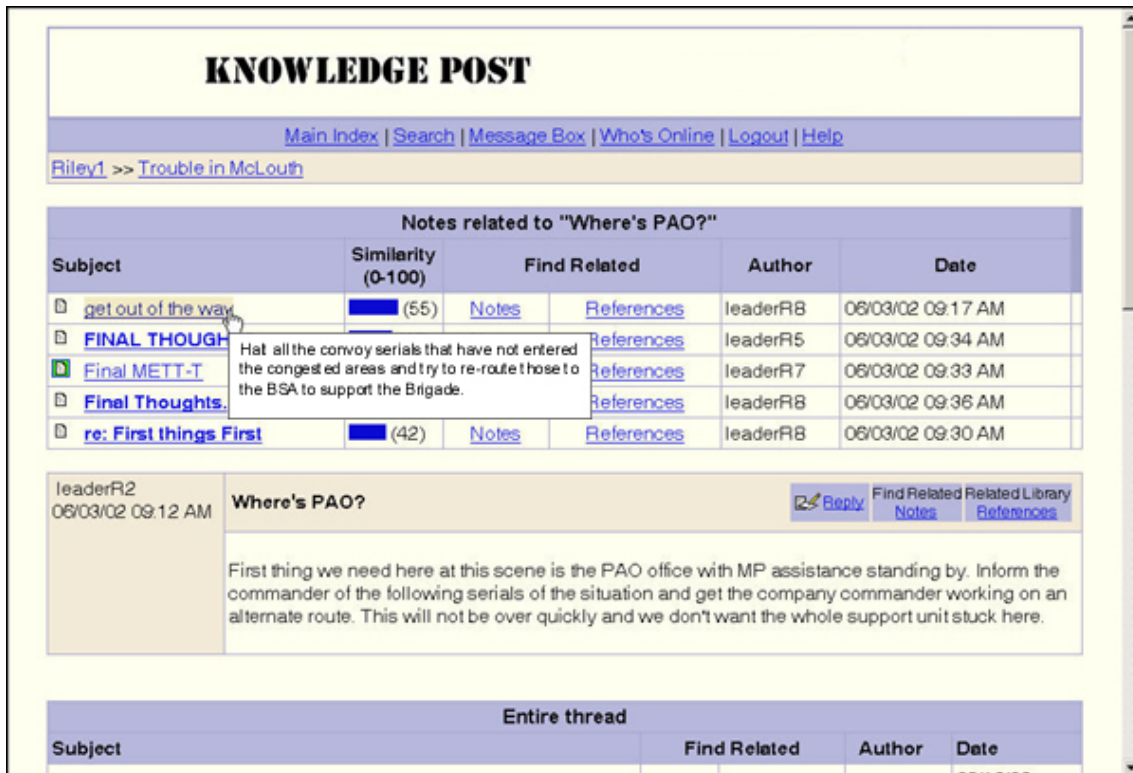
Rather than have respondents rate pre-generated alternatives as usual, after reading the scenario, we had officers write open-ended responses as to the course of action they would take. They were given instructions orally as part of their introduction to the threaded discussion environment.

Online Discussion Group Environment—Knowledge Post (KP)

Knowledge Post is a standard threaded discussion environment that has been enhanced with LSA. Its main features are as follows:

1. Users read scenarios or discussion prompts stored as notes within the discussion forum.
2. Users write notes in response to those scenarios or prompts.
3. Users can read and respond to the notes of other users.
4. Users are provided links to notes that are semantically similar to theirs or others.
5. Users can have their own responses evaluated by the *Intelligent Essay Assessor* (Foltz et al., 1999).

Features 1-3 are commonly available in other threaded discussion groups. Features 4 and 5 however are unique to *Knowledge Post* and rely on LSA.



KNOWLEDGE POST

[Main Index](#) | [Search](#) | [Message Box](#) | [Who's Online](#) | [Logout](#) | [Help](#)

[Riley1](#) >> [Trouble in McLouth](#)

Notes related to "Where's PAO?"				
Subject	Similarity (0-100)	Find Related	Author	Date
get out of the way	(55)	Notes References	leaderR8	06/03/02 09:17 AM
FINAL THOUGH		References	leaderR5	06/03/02 09:34 AM
Final METT-T		References	leaderR7	06/03/02 09:33 AM
Final Thoughts		References	leaderR8	06/03/02 09:36 AM
re: First things First	(42)	Notes References	leaderR8	06/03/02 09:30 AM

leaderR2
06/03/02 09:12 AM

Where's PAO?

[Reply](#) | [Find Related](#) | [Related Library](#) | [Notes](#) | [References](#)

First thing we need here at this scene is the PAO office with MP assistance standing by. Inform the commander of the following serials of the situation and get the company commander working on an alternate route. This will not be over quickly and we don't want the whole support unit stuck here.

Entire thread			
Subject	Find Related	Author	Date
			06/03/02

Figure 1. Notes semantically related to leaderR2's note entitled, "Where's PAO?". A PAO is a public affairs officer.

Semantically Related Notes: A Unique Feature of Knowledge Post

In many discussion boards, there are simply too many notes for a participant to read every one. LSA provides a remedy to this situation by allowing users to find contributions that are similar to their own. It thus provides a semantic path through the discussion board. The LSA enabled "Find Related Notes" function in Knowledge Post provides officers with a tool that allows them not only to read responses to their entries, but also read selectively other entries that are not only similar in meaning, but that may introduce related issues. Figure 1 shows the output of "Find Related Notes". The related notes are ordered in the display by their semantic similarity as shown as a number between 0 and 100 (cosine value multiplied by 100). Typically, similar notes have cosines that range from 0.5 to 0.8, overlapping the range of effective values found in Wolfe et al.

The semantic space that LSA creates depends critically on the text corpus over which it is computed. In Knowledge Post, this text corpus is very large, comprising about 12 million words from selected texts that form a representative sample of what a student finishing high school may have read. In addition, the content of several Army manuals, technical books, and descriptions of common Army scenarios are also

incorporated. In the past, LSA has shown a robust representation of the meaning of words and passages using similar spaces.

The assumptions underlying LSA propose that similarities and differences in the meanings of words can be largely induced from the contexts in which they appear, and that similarities and differences in the meanings of paragraphs can be induced from a combination of the constituent words. The order of these component words is not nearly as important as the particular selection of words. The meaning is induced not just from keywords, or local co-occurrence of words, but by the solution of many simultaneous equations. Each written contribution is given a position in a multidimensional LSA space based on the sum of the vectors of all its constituent word meanings. To determine the similarity between two contributions, we normally use the cosine of the angle between vectors in the semantic space as a similarity metric.

METHOD

Scenarios

Four Tacit Knowledge of Military Leadership scenarios were presented to officers within KP. A separate topic, on unit readiness, provided an open ended discussion of the officers' understanding of US Army readiness for combat, a central concern for all of them in their daily interactions.

Participants

Eight groups of officers at four Army installations participated in the experiment. Each installation sets aside one week a year during which its personnel are available for research, and the experiment was conducted during this period. Data were collected in small groups of 5 to 15 officers, where group members were all of the same rank. A total of 46 soldiers (11 LTs, 12 CPTs, 13 MAJs, and 10 LTCs) participated for the four platoon scenarios, while a different set of 47 Non Commissioned and Commissioned Officers discussed the Unit Readiness scenario.

Procedure

Each session lasted three hours or less. Copies of the instructions and additional assessment instruments were passed out prior to the experiment. Two introductory pages preceded the instrument. The first page described the purpose of the research and the role of the participant, with an informed consent form. The second page asked for optional information including rank, branch, and current duty position. Oral instructions were provided for logging on to the threaded discussion, and how to respond to each scenario. The scenarios were followed in the same order for each group.

RESULTS

Evaluating the quality of the online contributions.

A total of 199 notes was generated by the Officers for the four platoon scenarios. Each note was scored blind by four expert graders based on their best understanding of the ways to deal effectively with the issues in the scenario and also on the standardized answers available from the extensive administration of the multiple forced-choice 9-point scaling versions of the scenarios. Mean intercorrelations among the four raters of their grades over all 199 notes were significant ($p < .01$): 0.52, 0.40, 0.42, and 0.46. The correlations between any two graders ranged from a high of 0.72 to a low of 0.23.

For each contribution, the most similar notes were found using the cosine similarity metric. We compared the rated quality of each contribution and the rated quality of the nearest note to it, as well as computing the correlation between these two measures. The results are shown separately for each platoon scenario in Table 1.

All four scenarios demonstrated the Goldilocks Principle at work-- the quality of the near notes selected automatically by LSA was significantly higher than the quality of the original note used to search (stem note). At the same time, three of the four scenarios showed a significant correlation between the quality of the stem note and the quality of the nearest note selected by LSA.

Scenario	Mean Quality of Stem Note	Mean Quality of Near Notes	Correlation Between Stem and Near Note Quality
Platoon 1 (N = 47)	6.46	7.38**	0.42**
Platoon 2 (N = 57)	6.94	7.37**	0.28**
Platoon 3 (N = 54)	5.60	6.56**	0.25**
Platoon 4 (N = 41)	6.35	6.66**	0.01

** $p < .01$ (Column 3 significance is based on the difference between the Near Notes and Stem Notes)

Table 1. Average quality of original note and near notes and the correlation between the two.

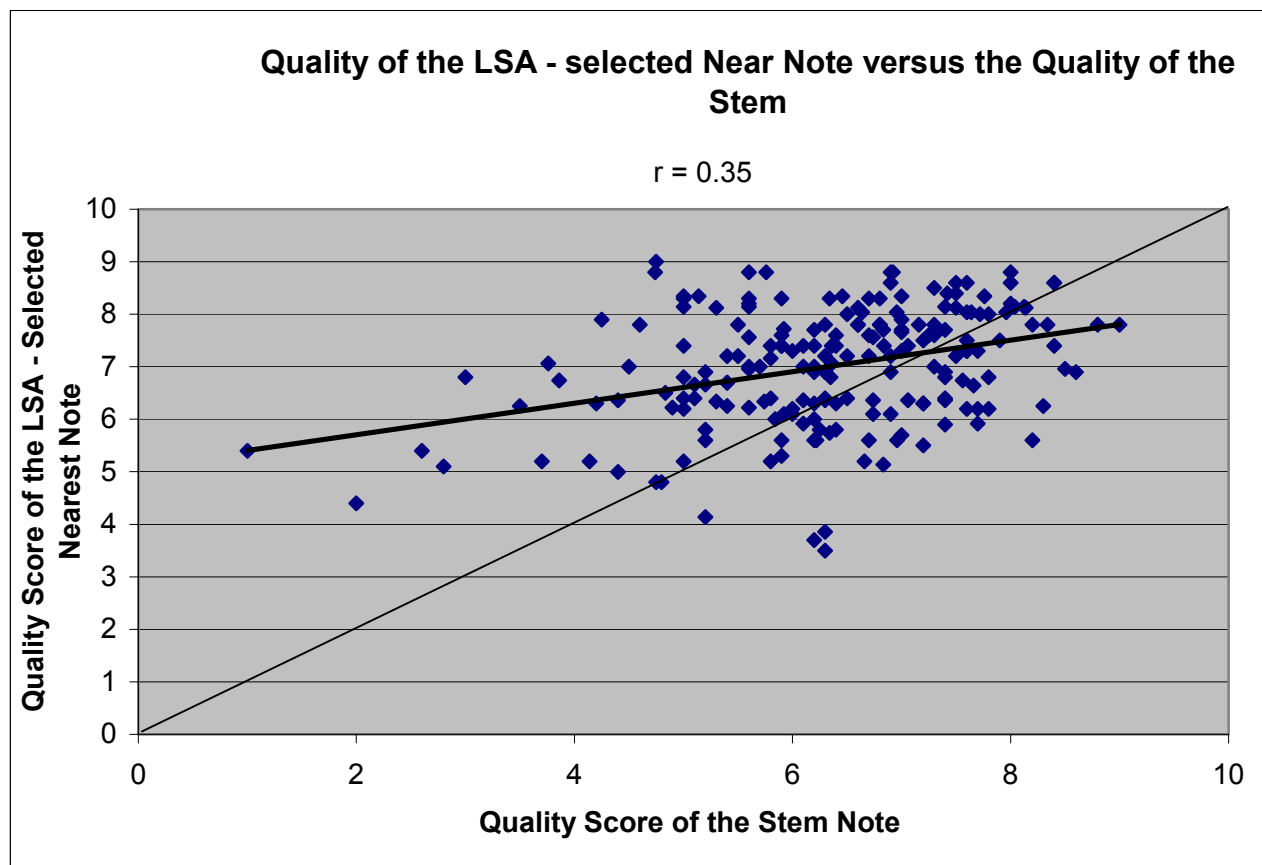


Figure 2. Correlation between the Stem Score and the LSA – selected nearest note score over all 199 notes in the platoon level scenarios.

The overall correlation between the quality of the Stem Note and the quality of the Nearest Note selected by LSA is moderate, but significant ($p < .01$) as shown in Figure 2. However, for the lower half of the Stem Notes, those with a quality score of less than 5, the Near Note is always of higher quality. For those Stem Notes with a quality score greater than 5, the majority of selected Near Notes is also higher in quality.

Figure 3 shows the distribution of the Stem and Near Notes. Of particular interest is the asymmetry of the Near Note distribution, which is markedly skewed to the higher end of the distribution. An exception to this pattern is the lowest bin, but there appears to be an explanation that highlights this exception. These low quality contributions were upon examination, usually irrelevant or extraneous and short comments, outside the knowledge domain; so LSA often linked them to each other but not to the higher quality contributions. These low quality comments are best seen as irrelevant to the issue of how to implement the Goldilocks Principle automatically in LSA.

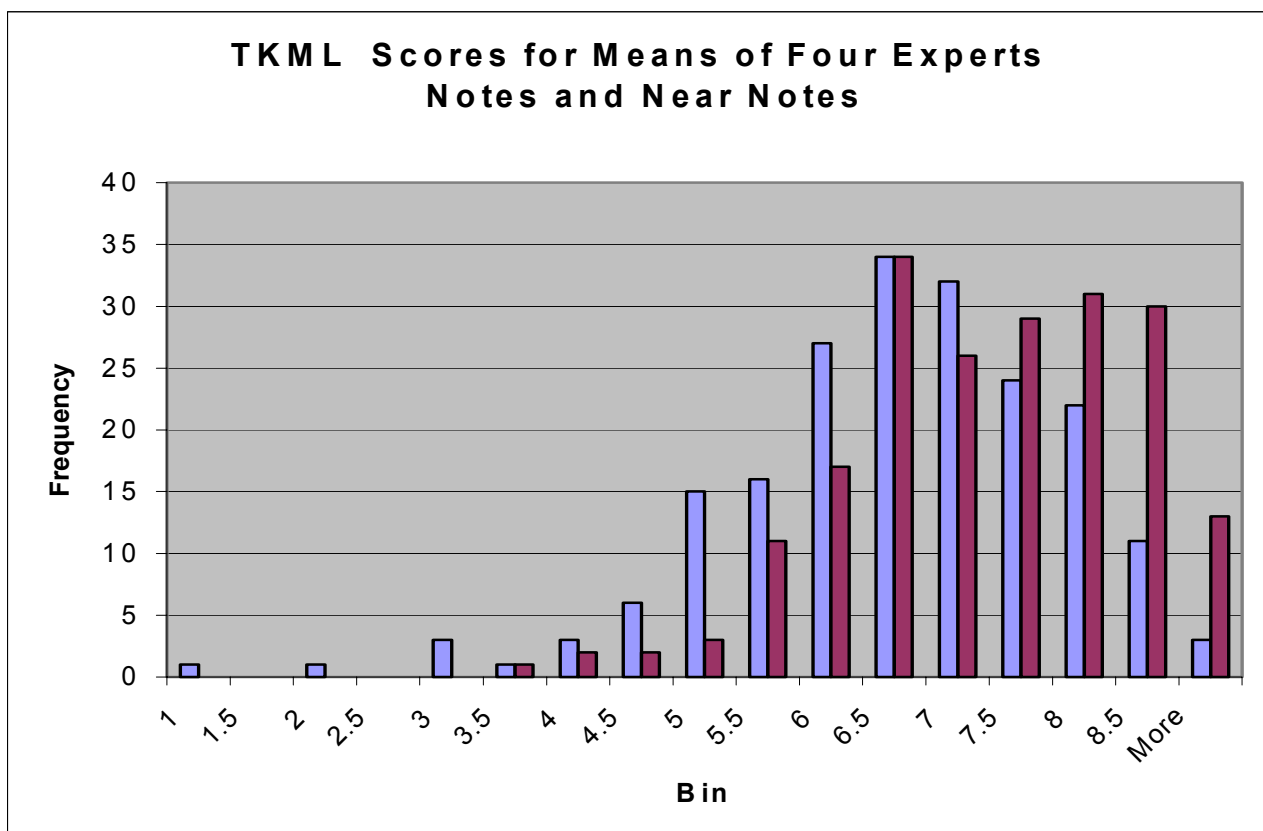
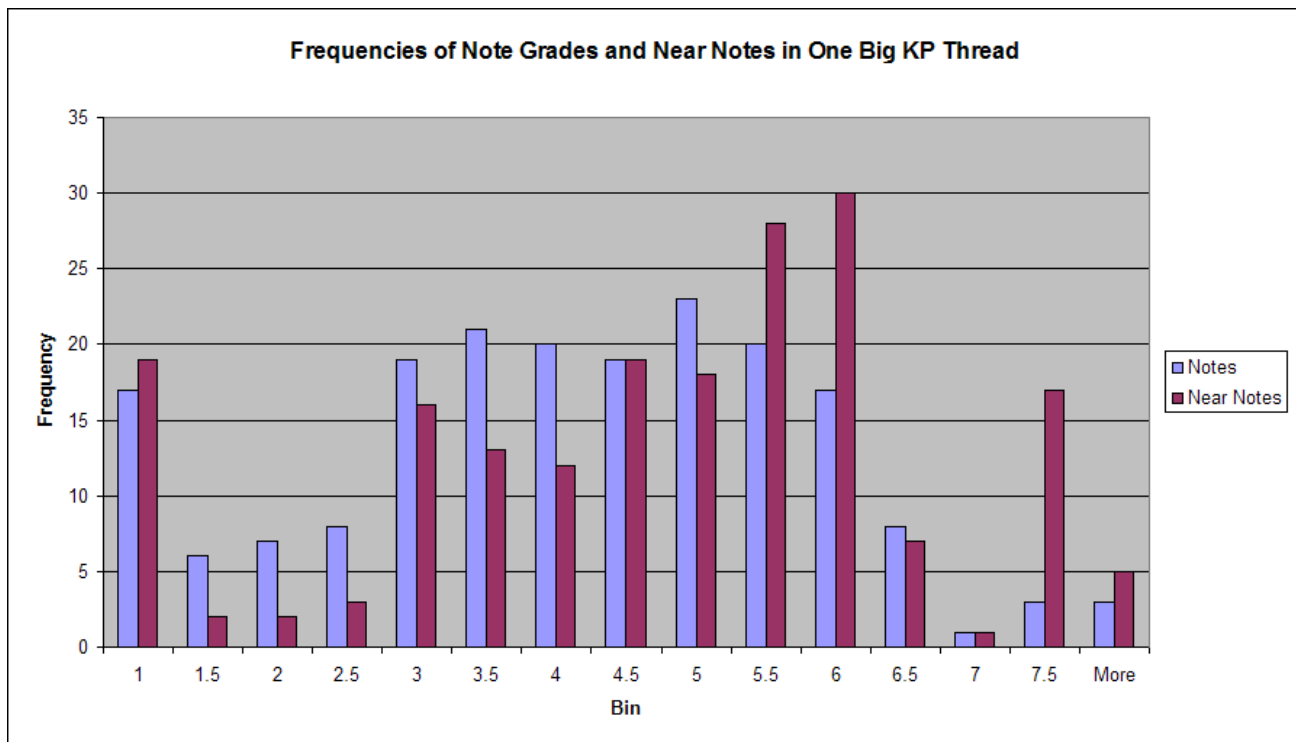


Figure 3. Distribution of mean Near Note grades and Stem Notes for the four Platoon scenarios. Stem notes are shown in blue; Near Notes in maroon.

Unit Readiness Results

The Unit Readiness responses were scored by two military expert graders using their best understanding of the components of readiness in the US Army. Mean intercorrelations between the two raters of their grades over all 192 notes were significant ($r = 0.89$; $p < .01$).

In all, the quality of the Near Notes selected automatically by LSA was significantly higher than the quality of the Stem Note initiating the search ($t = 5.2$; $p < .001$). At the same time there was a significant correlation between the quality of the Stem Note and the quality of the Nearest Note selected by LSA ($r = 0.56$; $p < .01$; 190 df).

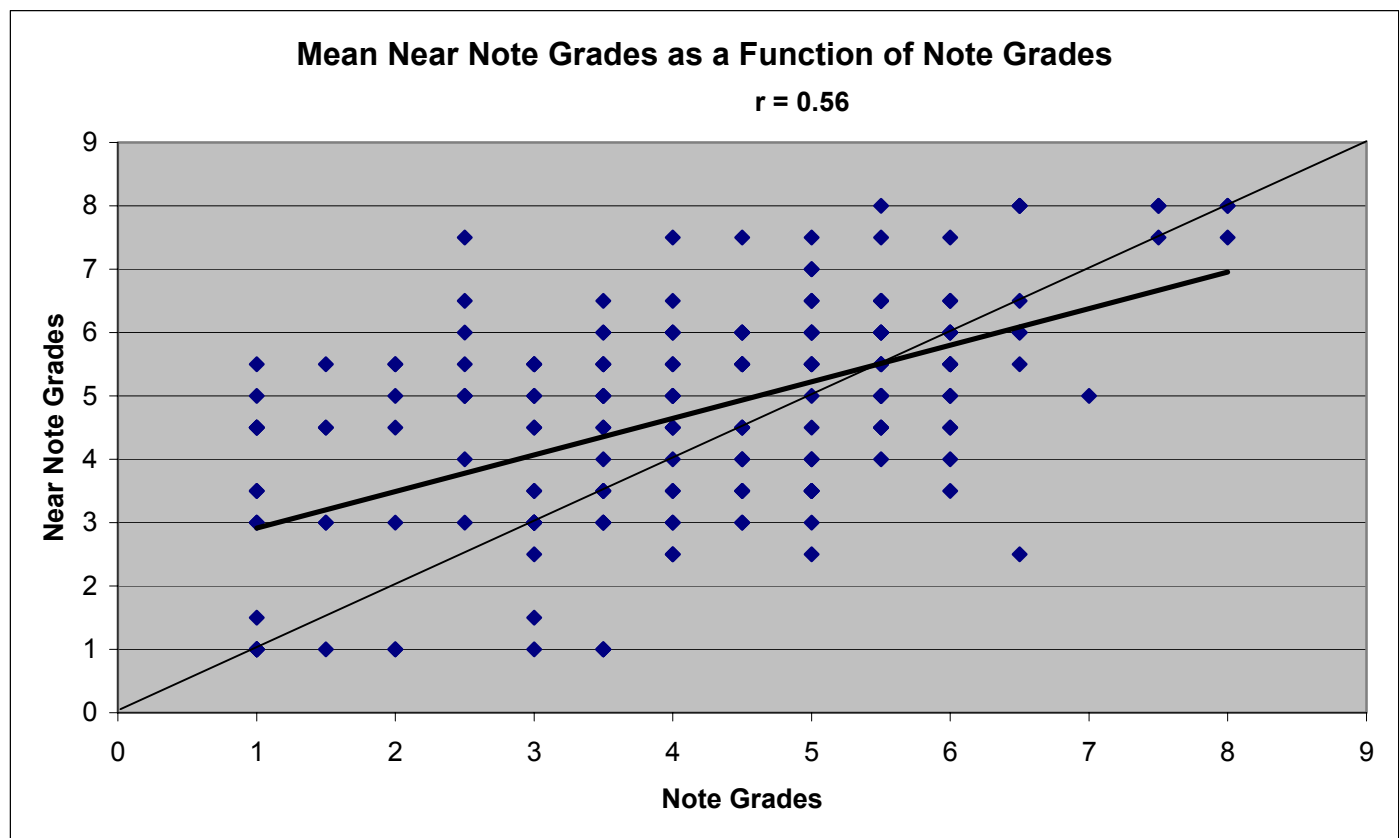


Figure 4. Mean Near Note grades as a function of the related note grades for Readiness Scores.

Comparing the distribution of the notes' grades against the related Near Note grades, the frequencies of the Near Note grades are relatively higher for the higher grades (cf. Figure 5), but the pattern of asymmetrical document-to-document cosines is not nearly as marked as it was for the platoon scenarios. Once again the very low quality notes appear to be selected too often. However, as in the TKML scenarios, these low quality notes are often asides and personal remarks that really should be excluded from the distribution, for the purposes of these analyses.

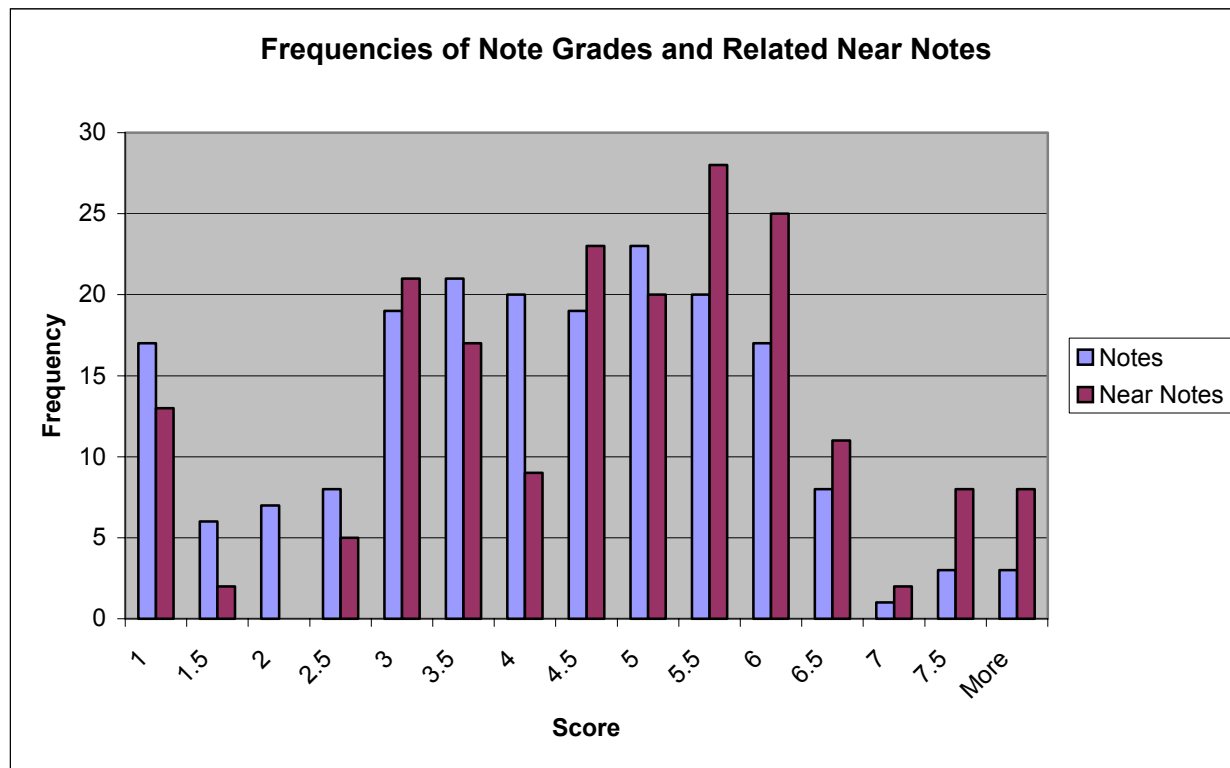


Figure 5. Frequencies of Note Grades compared to the frequencies of Near Note Grades for Readiness Scores.

ANALYSES

On average, in Latent Semantic Space a Near Note is of slightly higher quality than the corresponding Stem Note. But does *reading* a slightly better note lead to more effective learning as Wolfe et al. found? The evidence with regard to the latter question is probably affirmative, but wasn't directly tested in this study.

We have compared officers discussing these same scenarios either face-to-face or using our electronic discussion group (Lochbaum et al., 2002). After reading the scenarios about half of the officers discussed the scenario face-to-face and wrote their responses with pencil and paper, while the other half typed their "solutions" into the online discussion environment. The online group entered an initial response and then a final response after an online synchronous discussion.

Randomized responses were graded blind by two military experts for quality on a 1 to 9 scale with a reliability of 0.78. The results are shown in Figure 6. Those officers who used the online discussion group contributed much higher quality initial responses (shown as First Knowledge Post in the figure below) than the pencil and paper group. Additionally, lower ranking officers learned more using the online discussion group than did the face-to-face participants. How much of this effect can attributed to finding Near Notes while discussing the scenario is not known.

Officers were shown the Near Note facility and used it during the discussions, so some benefits may have accrued. However, this was only one of several factors that may have contributed to higher quality responses. For example, one of the oldest results from the work on electronic chat groups pioneered by Hiltz and Turoff (cf. Benbunan-Fich, Hiltz, & Turoff, 2003) is the greater equality of participation in the electronic medium over face-to-face groups. Anonymity certainly contributes to the increased participation—"nobody

knows you're a dog on the internet." In post discussion feedback we have collected on *Knowledge Post*, anonymity is always mentioned as a positive feature, producing richer, more honest communication.

Thus, several factors conspire to produce better discussion and learning in the electronic medium. One is natural desire of humans to communicate with each other. Another is the parallel nature of the discussion—members of an electronic discussion can contribute simultaneously, thereby making more effective use of the time available. This is not possible in face-to-face discussions. The simultaneity of input coupled with greater equality of participation, result in a richer set of ideas generated by a greater number of people. In face-to-face discussions only a few people contribute the bulk of the remarks—in small groups the most vocal two people do over 60% of the talking (Stephan & Mishler, 1952).

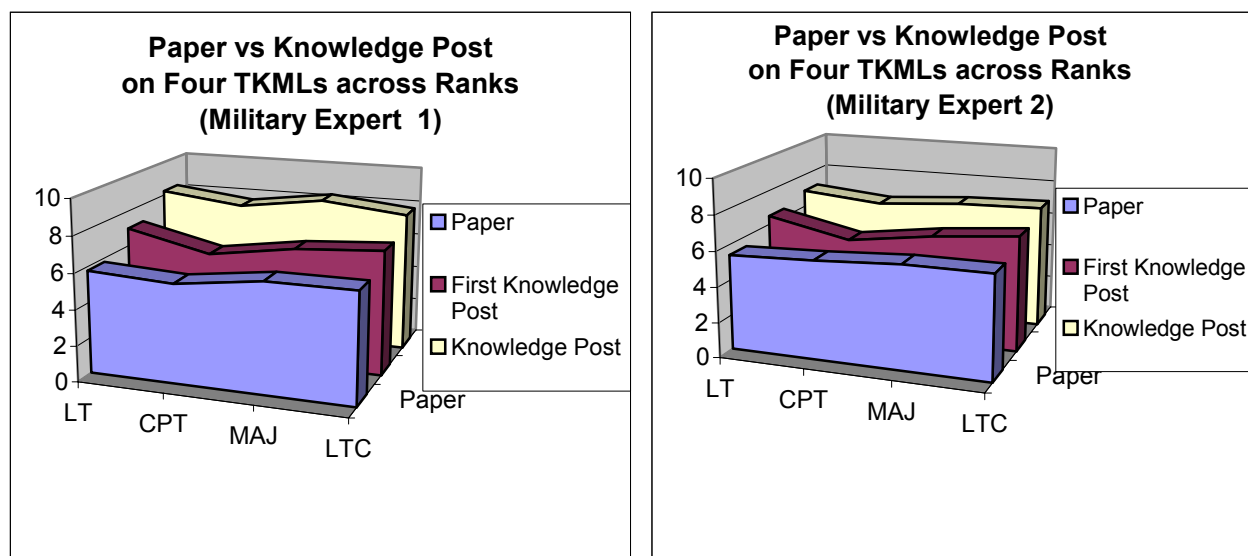


Figure 6. Comparisons of Paper and Pencil to Knowledge Post by Officer Level for the TKML scenarios for two Military Experts' grades (1-9 with 9 being the highest grade).

LSA and Goldilocks

How does a LSA representation of contributions automatically implement the "Goldilocks Principle" in which feedback and additional information are given at nearly the right level of difficulty—the next passage to read being somewhat more advanced than the original? An explanation depends on understanding the properties of the LSA space.

We computed the "centroid" of the high-dimensional vectors all the contributions and then asked whether it was more similar to high quality responses (best one-sixth) or more similar to low quality responses (worst one-sixth). The centroid was much closer to the best responses in terms of cosine similarity (mean cosine between the centroid and the best = 0.74 (SD=.07); mean cosine between the centroid and the worst = 0.60 (SD=.08)). Across all five scenarios the effect was significant ($p < .001$ or less in each case).

Regression to the Mean

Some sort of regression to the mean, either statistical or biased by LSA, can be ruled out as an alternative explanation of this phenomenon since the means of the near note distributions are all significantly higher

than the means of all the notes. From the scattergrams in Figures 2 and 4 it appears that the effect is somewhat stronger for the poorer contributions, but this may just be a ceiling effect for the better contributions, since there are a very limited number of even better contributions for the system to choose automatically.

Consensual Centroids of the Semantic Space

Another explanation of this phenomenon would be that the distributions of note to note cosine similarities are highly asymmetric, because there are more ways to be bad than good with poor contributions being more different from each other than good ones. Thus, low scores will be farther from the centroid and closer to better notes than poorer notes. This results in a near neighbor having a higher score on the average.

To expand on this analysis, as Tolstoy (1877) pointed out in the opening lines of *Anna Karenina*: "Happy families are all alike; every unhappy family is unhappy in its own way."

This Tolstoy principle might be applied to the consensual assessment of quality by LSA of "happy" responses. High quality responses tend to share many features and converge in meaning. Since the officers each have some expertise that they bring to bear on the problem scenario, the common consensual components tend to be those that offer the best solution to the problem scenario, and responses that include more of these components are in general more expert and of higher quality. As a result, the knowledge space of contributions in the threaded discussion tends to center on the consensually best answers, with worse responses scattered in the periphery of the space.

This explanation depends on the information captured by LSA's centroid of the semantic space of all contributions, and the distribution of responses in this semantic space. Better answers lie nearer the center of LSA's semantic space of these contributions, and these answers are best because they capture most of the consensual common components of experts. In the case of the Army officers, the centroid represents the parts of the solution that have been agreed on by several different people, without those components that are not generally agreed on. In this sense it is better than the average contribution. This effect has been seen in other, but not all, sets of essay scores examined.

We tested this notion directly with the four platoon Tacit Knowledge scenarios by examining the existing 9-point rating alternatives that have been normed for these particular scenarios (Hedlund et al., 2000). These alternatives capture succinctly the common recommendations that officers make about things to do under the circumstances. Some of these alternatives are seen as distinctly better than others. For example, for the scenario in which the new platoon leader failed Ranger School and has to command soldiers who have just returned from combat, the text of two relatively good and bad alternatives are shown below:

"Ask the members of the platoon to share their combat experience: Ask what they learned and how it can help the platoon."

(median 9 "extremely good" on a 9-point scale N=358)

"Announce right up front that you are in charge and the soldiers must accept this fact and treat you with appropriate respect."

(median = 3 "somewhat bad" N=358)

To examine whether the centroid contribution had more elements of good responses, we compared a note constructed of all good alternatives (median rating of at least 8), selected from the forced choice text versions of the TKML, with the centroid as well as with the best and worst one-sixth contributions for each of the platoon scenarios. The results are shown in Table 2. Low quality discussion contributions have significantly less similarity to the centroid than do high quality contributions ($t = 6.84$; 62 d.f.; $p < .01$). The

contributions that were combined from the text of good alternatives to the scenarios are closer to the high quality contributions than the low ($t = 4.38$; 62 d.f.; $p < .01$). The magnitude of these cosines is significantly less ($p < .001$) than with the real contributions. Evidently, the actual discussion contributions are more detailed than the distilled ones in the TKML.

Text Items	Proximity to Centroid	Proximity to Combined Good Rating Items
Low Quality Contributions (N=32)	0.60 ** (SD=.08)	0.19 ** (SD=.10)
High Quality Contributions (N=32)	0.74 (SD=.07)	0.30 (SD=.10)

Table 2. Cosine similarities (Proximity) to the centroid of high and low quality contributions, as well as similarities to the combined text for good rating, forced choice items across the four platoon scenarios used. Low versus High Quality mean differences are both significant **($p < .01$)

As a further test of the hypotheses, the authors created summaries of all the contributions for each of the rank groups ((LTCs, MAJs, CPTs, and LTs) trying to capture all of the best points that came up consensually and repeatedly in the discussion. Figure 7 provides a view of the distribution of all *readiness* comments in a space of cosine distances, with these author's summaries of each echelon's contributions (LTCs, MAJs, CPTs, and LTs) in orange clearly near the center of the distribution.

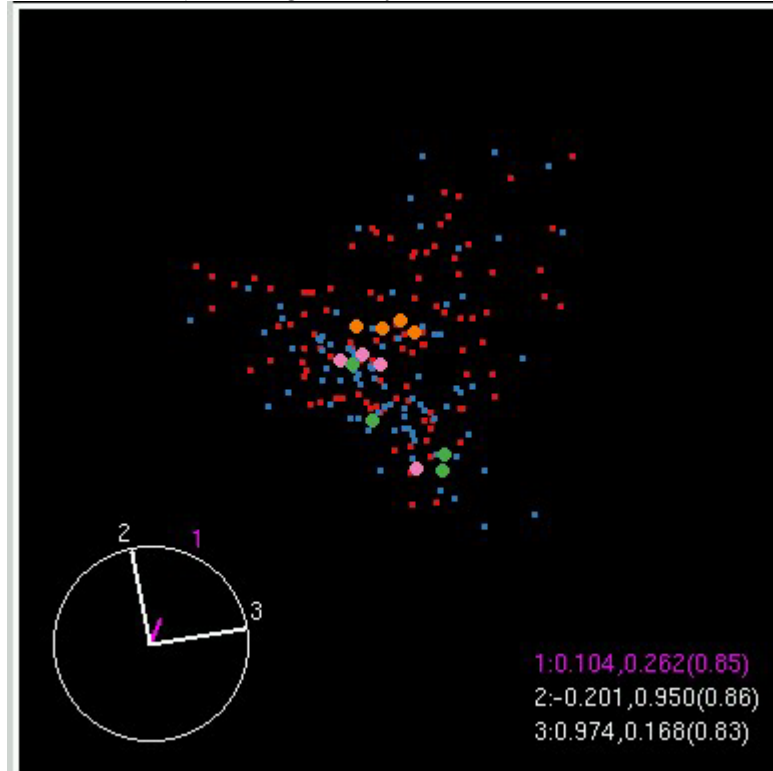


Figure 7. A multidimensional scaling of the cosine distance among readiness responses, with author – generated summaries apparently falling near the center of the space.

DISCUSSION

The Intrinsic Value of Consensus.

Another potential explanation is that notes nearer the center of the space better capture the consensus of the officers, and that the consensus is better not just because the opinions are in fact objectively better, but because consensus itself is valuable. Decision makers need to agree in order to act in concert and to motivate others. The better leaders have learned how to reach consensus more quickly and effectively.

A perhaps somewhat less interesting version of this explanation is that the raters have reached consensus on what is good while reading the notes, and the participants, also being officers, have tended to reach the same consensus. This is less interesting only in the sense that consensus reached by long real-world experience and interaction in leadership situations would probably be more valid than consensus reached through reading discussion notes.

However, the most interesting explanation resides in the possibility that consensus opinions on difficult issues really are better, and somehow closer to the truth, even when there is no absolute or objective standard to make that assessment. Somehow, out of experience, these officers are intuiting parts of better approaches to dealing with complex problems, and the sum of these approaches constitutes something like the best humans can do in these situations. Thus each officer's contribution can be seen as a correlation with this best approach; and their intercorrelations dependent on this correlation with the best approach.

Using LSA to Implement ZPD

The finding that LSA represents a semantic space of contributions to a threaded discussion in which the centroid appears to be composed of better components determined consensually from the space of all answers provided in the discussion provides some insight into how the ZPD can be applied with LSA. By offering a Near Note that is slightly better than each officer's own response, LSA could offer additional (and better) components that are not in the officer's own response, while maintaining some commonality with his or her own response. The officer then has the opportunity of expanding his or her own views to try to encompass the new components, hopefully improving his or her understanding of the issues at the same time. LSA is also sensitive to word grade or difficulty level, so it seems reasonable to assume that LSA or other statistical measure can be used to maintain grade level of vocabulary while introducing more consensual components.

Although we have not tested this hypothesis explicitly, we consider it an interesting direction for further research and we plan to examine other means for testing it empirically.

The way in which LSA models consensus provides an automatic method for selecting the next best piece of material for a student to learn, making this an important contribution to tutoring. Further research is needed to determine more precisely how effective this implementation is in actual learning situations. From the scattergrams, it appears that poor responses elicit near notes that are much better than they are, and this distance may be too great to implement the Goldilocks Principle optimally. For these notes, it may be more effective to have the system automatically select a second or third order near note. Fine tuning of these effects should be relatively easy to implement automatically. However, the evidence from these five case examples is strong that on the whole effective tutoring is occurring automatically in the current system, and provides a first-order implementation of the Goldilocks Principle within the Zone of Proximal Development proposed by Vygotsky (1978).

References

- Anderson, J. R. (2002) Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, Vol 26(1), Jan-Feb 2002. pp. 85-112. Journal URL: <http://www.elsevier.com/inca/publications/store/6/2/0/1/9/4/>
- Bloom, B. S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.
- Bruner, J. S. (1966). *Toward a theory of instruction*. New York: Norton.
- Burton, R. R., & Brown, J. S. (1979). An investigation of computer coaching for informal learning activities. *International Journal of Man-Machine Studies*, 11, 5-24.
- Foltz, P. W., Laham, D., Landauer T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Education Journal*. Available online at: (<http://imej.wfu.edu/articles/1999/2/04/index.asp>).
- Hedlund, J., Sternberg, R. J., & Psotka, J. (2000). Tacit knowledge for military leadership: Seeking insight into the acquisition and use of practical knowledge (Tech. Rep. No. ARI TR 1105). Alexandria, VA: U.S. Army Research Institute.
- Benbunan-Fich, R., Hiltz, S. R., & Turoff, M. (2003) *Decision Support Systems*, Vol 34(4), Mar 2003. pp. 457-469.
- Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In Ross, B. H (Ed.) *The Psychology of Learning and Motivation*, Vol 41, 43-84.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, T. K., Foltz, P. & Laham, D. (1998). An introduction to Latent Semantic Analysis, *Discourse Processes*. 25, 259-284.
- Lochbaum, K., Streeter, L, & Psotka, J. (2002). Exploiting technology to harness the power of peers. *Interservice/Industry Training, Simulation and Education Conference*, Orlando, FL, December 2-5, 2002.
- Palincsar, A. S., & Brown, A. L. (1984). The reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1, 117-175.
- Psotka, J., Massey, L. D, & Mutter, S. A. (1988). *Intelligent tutoring systems: Lessons learned*. Erlbaum, 552 pp.
- Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, Vol 25(2-3).
- Stephan, F. F. & Mishler, E. G. (1952). The distribution of participation in small groups: an exponential approximation. *American Sociological Review*, 17, 1952. pp. 598-608.
- Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.
- Tolstoy Leo (1877) *Anna Karenina*. Great Literature Online. [Http://www.underthesun.cc/](http://www.underthesun.cc/)

Vygotsky, L. (1978). Mind in Society: The Development of Higher Psychological Processes. Cambridge, MAL: Harvard University Press.

Wolfe, Michael B. W.; Schreiner, M. E.; Rehder, R.;(1998) Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, Vol 25(2-3), 1998. pp. 309-336.